*Jörg Faust*

# Rigorose Wirkungsevaluierung – Genese, Debatte und Nutzung in der Entwicklungszusammenarbeit

**Zusammenfassung**
Der Aufstieg rigoroser Wirkungsevaluierungen (RIE) in der internationalen Entwicklungszusammenarbeit (EZ) geht einher mit einer anhaltenden Kontroverse um experimentelle und quasi-experimentelle Methoden. Vor diesem Hintergrund und auf Basis eines empirisch-analytischen Wissenschaftsverständnis leistet der vorliegende Aufsatz Beiträge zur Erklärung der Genese von RIE im Politikfeld, zu einer konstruktiven Auseinandersetzung mit den Vorteilen und Begrenzungen von RIE aus evaluatorischer Perspektive sowie zur Analyse ihrer Anwendung und Nutzung in der Praxis der internationalen EZ. Im Ergebnis erfolgt ein Plädoyer für eine breite Anwendung und Nutzung von RIE in der EZ, ohne diese aufgrund ihrer Begrenzungen jedoch zum Königsweg der Evaluierung oder gar der empirischen Sozialforschung zu erheben. Vielmehr sollte die Anwendung gegenstandsangemessen, theoriebasiert und um qualitative Methoden ergänzt erfolgen und andere Evaluationsverfahren dort eingesetzt werden, wo RIE nicht geeignet sind.

*Schlagworte*: Evaluation, Entwicklungszusammenarbeit, Wirkungsevaluierung, Experimente, Methoden

**Abstract**
*Rigorous Impact Evaluation in Development Cooperation: Origins, Debates and Use*
The rise of rigorous impact evaluation (RIE) in international development cooperation has been accompanied by a controversy around experimental and quasi-experimental approaches. Against this backdrop and from an empirical-analytical perspective, this paper contributes to the ongoing debate in three ways: first, it explains the emergence of RIE in foreign aid from the particularities of the aid effectiveness debate; second, it contributes to an enlightened debate by sketching the advantages and limitations of RIE from an evaluation perspective. Finally, it sketches the current use of RIE in the practical realm of development cooperation. Overall, the paper argues in favour of a broader and more systematic use of RIE in development cooperation without disregarding the methodological, empirical and practical limitations of the approach.

*Keywords*: Impact Evaluation, Development Cooperation, Randomized Control Trials, Methods

## 1 Einleitung[1]

Mit der Vergabe des Alfred-Nobel-Gedächtnispreises für Wirtschaftswissenschaften 2019 an Esther Duflo, Abhijit Banerjee und Michael Kremer hat die Bedeutung von Feldexperimenten und quasi-experimentellen Designs in der Entwicklungsforschung

bzw. den damit befassten Sozialwissenschaften weiter zugenommen. Von der in den letzten Jahren stark angestiegenen Anzahl der Vertreter*innen experimenteller Feldforschung wird diese nicht selten als empirischer Königsweg der Entwicklungsforschung aufgefasst und deren weitere Verbreitung eingefordert (Banerjee & Duflo, 2011). Gleichzeitig formierte sich insbesondere innerhalb der qualitativ arbeitenden Sozialforschung Widerstand gegen den vermeintlichen Goldstandard und seinen quantitativ-erklärenden Zuschnitt zu Lasten einer eher qualitativ-verstehenden Wissenschaftstradition. Doch auch quantitativ arbeitende, einem empirisch-analytischen Wissenschaftsverständnis verpflichtete Ökonom*innen wandten sich gegen eine unkritische Verbreitung der experimentellen Forschung, so etwa der Nobelpreisträger Angus Deaton (u. a. Deaton, 2010; Pritchett & Sandefur, 2015).

Ihren Niederschlag in der angewandten Forschung fand diese Debatte insbesondere in der Evaluierung internationaler entwicklungspolitischer Maßnahmen. Auch hier wurde seit den 2000er Jahren eine Zunahme rigoroser Wirkungsevaluierungen eingefordert und auch – wenngleich keinesfalls systematisch – durchgeführt (u. a. Centre for Global Development, 2006). Als rigorose Wirkungsevaluierungen (*rigorous impact evaluation* - RIE) werden im Folgenden Evaluierungen begriffen, die auf die Identifikation der Wirkungen einer Maßnahme auf Zielgruppenebene fokussiert sind und sich hierbei experimenteller *oder* quasi-experimenteller Methoden bedienen. Dabei steht die kausale Zuordnung möglicher Veränderungen bei der Zielgruppe (z. B. Individuen, Haushalte, Schulen, Wirtschaftsakteure) zu einer oder mehrerer zuvor erfolgten Maßnahme im Mittelpunkt.

Im Kern besteht der methodische Ansatz darin, einen Vergleich einer Interventionsgruppe mit einer möglichst ähnlichen Kontroll- oder Vergleichsgruppe durchzuführen. Ein solcher Vergleich ermöglicht „dann eine empirisch fundierte Aussage darüber, wie sich die Zielgruppe einer Intervention sowohl mit als auch ohne Durchführung der entwicklungspolitischen Maßnahme entwickelt hätte“ (Bruder, Faust & Krämer 2019, S. 1).[2] RIE können sowohl am Ende eines Programms als summative Evaluierung als auch im Sinne einer wirkungsorientierten Begleitforschung oder formativen Evaluierung bereits im Verlauf der Umsetzung von Maßnahmen eingesetzt werden (u. a. Pritchett & Sandefur, 2015; Funk, Groß, Leininger & Schiller, 2019).

Rigorose Wirkungsevaluierungen sind gemäß ihrer Befürworter*innen von hohem Nutzen, können sie doch bei richtiger Anwendung mit einer im Vergleich zu anderen Methoden deutlich höheren Sicherheit die Wirkungen von entwicklungspolitischen Maßnahmen bzw. deren Ausbleiben identifizieren. Entsprechend groß könne ihr Beitrag für eine evidenzbasierte Politikgestaltung in einem Politikfeld sein, das immer noch einer intensiven und nicht selten ideologischen Debatte zwischen Verfechtern externer Hilfe und deren Gegnern ausgesetzt ist. Eine auf RIE basierende Evidenzagenda solle dabei weit über einzelne RIE hinausreichen (White, 2019). Eine solche Evidenzagenda enthält auch die regelgebundene Aggregation von RIE in systematischen Reviews, um die externe Validität zu erhöhen und hierdurch projektübergreifendes Lernen zu ermöglichen. Ebenso sind Evidenzkarten Teil dieser Agenda, die mittels einer systematisch erfolgenden Sammlung und Typologisierung von RIE nach thematischen Fragestellungen das bereits vorhandene Wissen sowie Evidenzlücken gleichsam kartographieren. Schließlich soll die Nutzbarkeit existierender Erkenntnisse über die Einrichtung von Evidenz-Portalen verbessert werden, die das vorhandene Wissen der globalen Gemeinschaft einfach und zweckdienlich zur Verfügung stellen (White, 2019).

Demgegenüber positionierten sich die Kritiker*innen von RIE mit einer Reihe von Argumenten, welche die Anwendung von RIE zumindest relativieren. Zu diesen Argumenten zählen unter anderem die begrenzte thematische Reichweite bzw. Anwendbarkeit von RIE. Zudem sei die externe Validität von RIE sehr gering. Auch basierten viele RIE auf einer mangelnden theoretischen Fundierung, vernachlässigten die Analyse von Kausalmechanismen, seien unverhältnismäßig teuer und aus ethischen Gründen abzulehnen. Schließlich seien RIE politisch neoliberal aufgeladen und es bestünde oftmals eine Verzerrung zu Gunsten der Kontrollfunktion von Evaluierung, was deren Nützlichkeit für das Lernen in der Praxis der Entwicklungszusammenarbeit (EZ) begrenze.[3]

Angesichts des skizzierten Aufstiegs rigoroser Wirkungsevaluierungen in der EZ und der anhaltend kontroversen Debatte, leistet der vorliegende Aufsatz auf Basis eines empirisch-analytischen Wissenschaftsverständnis Beiträge 1) zur Erklärung der Genese von RIE im Politikfeld, 2) zu einer konstruktiven Auseinandersetzung mit den Vorteilen und Begrenzungen von RIE aus evaluatorischer Perspektive sowie 3) zur Analyse deren Anwendung und Nutzung in der Praxis der internationalen EZ.

Der Beitrag gliedert sich wie folgt: Nach der Einleitung wird der Aufstieg rigoroser Wirkungsevaluierungen aus der Kombination eines funktional abgeleiteten Bedarfs an Evaluierung und drei Katalysatoren einer intensiv geführten Wirksamkeitsdebatte im Politikfeld (Kapitel 2) begründet. In der Auseinandersetzung mit den Vorteilen und Begrenzungen plädiert dieser Beitrag zweitens für eine breitere Anwendung und Nutzung von RIE im Politikfeld und eine Stärkung der Evidenzagenda, ohne diese jedoch zum Goldstandard zu erheben. Vielmehr sollte die Anwendung von RIE gegenstandsangemessen, theoriebasiert und um qualitative Methoden ergänzt erfolgen. Dies öffnet den Raum für andere Evaluationsverfahren in einer Vielzahl von Anwendungsbereichen, in denen RIE nicht durchgeführt werden können bzw. nicht zweckdienlich für ein praxisgetriebenes Erkenntnisinteresse sind (Kapitel 3). Drittens schließlich verweist der Beitrag auf die Heterogenität bzw. Uneinheitlichkeit und Ungleichzeitigkeit der Anwendung, Verbreitung und Nutzung von RIE in der internationalen EZ und problematisiert die Fragmentierung der Gebergemeinschaft bzw. deren Wissensmanagement in diesem Feld (Kapitel 4).

## 2 Zur Genese rigoroser Wirkungsevaluierung in der EZ

Die Diskussion um Relevanz, Angemessenheit und praktische Verankerung rigoroser Wirkungsevaluierung in der EZ ist eng mit der breiter geführten Wirksamkeitsdebatte im Politikfeld verbunden. Um diese Verknüpfung mit der hier im Zentrum stehenden Evaluationsperspektive angemessen herauszuarbeiten, soll zunächst aus funktionaler Sicht der strukturell hohe Bedarf an Evaluation in der EZ begründet werden. Daraufhin wird argumentiert, dass dieser Bedarf an Evaluation durch drei Faktoren aus der breiteren Wirksamkeitsdebatte verstärkt wurde. Das Zusammenspiel des strukturellen Evaluationsbedarfs und der drei Katalysatoren haben im Politikfeld die Forderung nach bzw. Durchführung von rigorosen Wirkungsevaluierungen begünstigt, die in den letzten Jahren zentraler Bestandteil einer breiteren Agenda zur Nutzung rigoroser Evidenz im Politikfeld geworden sind.

## 2.1 Die strukturell hohe Nachfrage nach Evaluation in der EZ

Die Analyse der potenziellen Wirkungen in der EZ und deren direkte Anbindung an die Durchführungspraxis und politische Entscheidungsfindung mittels des Instrumentariums der Evaluierung hat eine lange Tradition. In nur wenigen anderen Politikfeldern wurde in den letzten vier Jahrzehnten so umfassend evaluiert wie in der EZ. Vielfach wurden hierbei auch Strukturen, Organisationen und Prozesse in den Blick genommen, so dass Evaluation nicht nur die Wirkungsanalyse zum Gegenstand hat. Neben der schieren Anzahl von Evaluierungen ist zudem die Institutionalisierung der Evaluation bemerkenswert. Diese schlägt sich erstens in Evaluationsreferaten, Stabsstellen, Kommissionen oder gar eigenständigen Instituten auf bilateraler Ebene wie auch in multilateralen Organisationen nieder. Zweitens ist die Institutionalisierung der Evaluation durch ein hohes Maß an (internationaler) Standardsetzung und organisationsspezifischer Regulationsdichte gekennzeichnet, was etwa Prinzipien, Kriterien und Durchführungspraktiken der Evaluierung anbelangt (im Überblick OECD, 2016; für Deutschland siehe u. a. Meyer, Bär, Faust, Jan, Silvestrini & Wein, 2019).

Dieser ausgeprägte Institutionalisierungsgrad und die hohe Evaluationsdichte in der EZ lassen sich aus einer funktionalen Perspektive begründen. Als spezifische Form angewandter Forschung soll Evaluierung drei miteinander verschränkte Funktionen erfüllen (Stockmann, 2004, S. 3). Mittels der Erkenntnisfunktion ist Evaluation auf die Schaffung praxisrelevanten, nützlichen Wissens ausgerichtet. Das generierte Wissen soll sodann über die Lernfunktion von den jeweils Verantwortlichen zur Verbesserung von Projekten, Programmen und Strategien eingesetzt werden. Gleichzeitig hat Evaluierung eine wichtige Kontrollfunktion, da ihre Befunde der Rechenschaftslegung über die Verwendung öffentlicher Mittel dienen. In Kombination können Erkenntnis-, Lern-, und Kontrollfunktion schließlich zur Legitimität eines Politikfeldes beitragen, wenn sie über die Bereitstellung der ermittelten Erkenntnisse für Lernen und Rechenschaftslegung sowohl Output- wie auch Inputlegitimität erhöhen.

Erkenntnis-, Lern- und Kontrollfunktion begründen den strukturell hohen Evaluierungsbedarf in der EZ. Die Erkenntnisfunktion von Evaluierung trifft auf ein hohes Maß an Unkenntnis bzw. Unsicherheit als einem Strukturmerkmal des Politikfeldes. Maßnahmen der EZ können als externe Interventionen in „fernen“ Gesellschaften begriffen werden. Diese Interventionen bzw. die sie durchführenden oder veranlassenden Akteure der EZ sind gemeinhin mit einem hohen Maß an Unsicherheit konfrontiert. Ob Brunnenbau, humanitäre Hilfe, Fortbildungen im Gesundheitssektor, Reformberatung auf Regierungsebene oder Korruptionsbekämpfung: Die Kenntnisse über die vor Ort existierenden Strukturen und Prozesse, auf die die Entwicklungsmaßnahmen treffen, sind vergleichsweise gering und zudem sind diese Strukturen und Prozesse oftmals durch ein hohes Maß an Instabilität gekennzeichnet. Entsprechend hoch ist der Bedarf an verlässlicher Evidenz über die Auswirkungen entwicklungspolitischer Maßnahmen und die diesen Wirkungen zugrundeliegenden kausalen Mechanismen.

Zudem sind Entwicklungsmaßnahmen gemeinhin keine einmaligen Vorgänge. Maßnahmen werden oft über etliche Jahre in mehreren Phasen fortgesetzt und – zunehmend kontextspezifisch angepasst – an anderen Orten repliziert. Daher gilt es, Erfahrungen und Erkenntnisse aus der Zusammenarbeit in einem Land oder in einer Region aufzubereiten, um diese in die Implementierung neuer Maßnahmen in anderen Kontexten einfließen zu lassen. Die umfassende geografische Reichweite der EZ und

die Notwendigkeit erfahrungsbasierter Anpassungen über Zeit, Sektoren und räumliche Grenzen hinweg machen Lernen und Wissensmanagement daher zu einem besonders wichtigen Erfolgsfaktor im Politikfeld. Entsprechend groß ist das Potenzial, Evaluation in der EZ als Lerninstrument in der eigenen Organisation, aber auch darüber hinaus organisationsübergreifend einzusetzen.

Schließlich ist die EZ durch ein strukturelles Kontrollproblem gekennzeichnet. Hierbei treten zwischen den unterschiedlichen Akteuren der entwicklungspolitischen „Wertschöpfungskette" ausgeprägte Prinzipal-Agenten-Probleme sowie eine fehlende Feedback-Schleife zwischen den Financiers der Zusammenarbeit und deren Endbegünstigten auf (u. a. Martens, Mummert & Murrell, 2002). Die Steuerzahlenden, aber oft auch deren parlamentarische Vertreter*innen verfügen gemeinhin über deutlich weniger Informationen und Kenntnisse über das breite Einsatzspektrum und die potenziell mannigfaltigen Auswirkungen von Entwicklungshilfe als über vergleichbare innenpolitische Sektorpolitiken wie Bildungs- und Gesundheitspolitik oder innere Sicherheit.[4] Gleichzeitig sind die eigentlichen Zielgruppen der EZ - marginalisierte und/oder bedürftige Bevölkerungsgruppen in Entwicklungsländern - kaum dazu in der Lage, ihre Erfahrungen systematisch zu bündeln und an die Steuerzahlenden oder deren Vertreter*innen rückzukoppeln. Zwischen den Steuerzahlenden und den Zielgruppen sind jedoch eine Vielzahl von besser organisierten und informierten Akteuren positioniert, die nicht ausschließlich die Verwirklichung originär entwicklungspolitischer Ziele verfolgen: Nationale Geberregierungen mit außenpolitischen und außenwirtschaftlichen Interessen; bilaterale und multilaterale Durchführungsorganisationen sowie auch zivilgesellschaftliche Organisationen mit bürokratischen Sonderinteressen und schließlich Staatsorgane in Empfängerländern, die vielfach nicht sonderlich gemeinwohlorientiert handeln (Martens, Mummert & Murrell, 2002; Easterly, 2002; Faust & Michaelowa, 2013). Insgesamt ist die EZ somit ein Politikfeld, in dem ein hoher Bedarf für unabhängige Evaluierung zur Stärkung von Transparenz und Rechenschaftslegung existiert.

## 2.2 Katalysatoren der Wirksamkeitsdebatte

Der zuvor funktionalistisch begründete Bedarf für ein überproportional hohes Maß an Evaluierung in der EZ ist in den letzten beiden Dekaden im Zuge einer kritischen Auseinandersetzung um die Wirkungen der EZ durch drei Faktoren verstärkt worden. Alle drei Faktoren haben hierbei gleichsam als Katalysatoren für die zunehmende Bedeutung von RIE fungiert.

*Ordnungspolitische Kritik*: Ein erster Katalysator ist die ordnungspolitisch begründete Skepsis gegenüber dem Politikfeld bzw. den hierin wirkenden Akteuren aus Staat und nationalen, internationalen und zivilgesellschaftlichen Entwicklungsorganisationen. Bereits in den frühen 1960er Jahren wurde „Entwicklungshilfe" als ein außenpolitisches Instrument erachtet, das – mit Ausnahme humanitärer Hilfe – in den Dienst außenpolitischer Interessen der Geber gestellt werden solle (Morgenthau, 1962). Seit den 1990er Jahren ist die Zahl an Allokationsanalysen gestiegen, die empirisch belegen, dass die Allokation von Entwicklungsgeldern auch ökonomischen und sicherheitspolitischen Motiven der Geberstaaten bzw. multilateralen Organisationen folgt (u. a. Dreher, Sturm & Vreeland, 2009; mit Blick auf die deutsche EZ u. a. Faust & Ziaja, 2012). Unabhängig davon, ob diese Berücksichtigung außenwirtschaftlicher oder sicherheits-

politischer Motive als legitim erachtet wird, kann deren Berücksichtigung negative Auswirkungen auf die Wirksamkeit entwicklungspolitischer Zielsetzungen haben. Denn die Ressourcen werden dann nicht mehr mit der gleichen Konsequenz in diejenigen Länder bzw. Projekte und Programme fließen, deren Aussicht auf entwicklungspolitische Wirksamkeit am höchsten oder zumindest aus entwicklungspolitischen Gründen angemessen ist (Dreher, Fuchs, Lang & Langlotz, 2018).

Neben der der Wirksamkeit abträglichen Verquickung von außenpolitischen mit entwicklungspolitischen Motiven bei der Vergabe, fokussiert ein weiterer ordnungspolitischer Kritikpunkt auf die Planungseuphorie einer über Jahrzehnte angewachsenen und gleichzeitig fragmentierten Entwicklungsbürokratie auf nationaler wie multilateraler Ebene (Easterly, 2002; 2006). Die Kritik lautet dabei, dass fehlerhafte Anreizstrukturen innerhalb dieser Organisationen den Fokus auf die Verausgabung von Mitteln, Projektproliferation und Abgrenzung gegenüber anderen Organisationen begünstige. Hingegen fehle eine marktkonforme Wettbewerbsstruktur, die wirksame Maßnahmen belohne (Easterly, 2002). Dadurch seien Mittelbeschaffung und Mittelverausgabung zentrale Erfolgsgrößen der Entwicklungsorganisationen geworden, nicht jedoch die Wirksamkeit der implementierten Maßnahmen; ein erst allmählich im Wandel begriffenes Strukturmerkmal der internationalen EZ.

*Miko-Makro-Paradoxon*: Komplementär zu der oben genannten ordnungspolitischen Kritik besteht der zweite Katalysator für die zunehmende Bedeutung von RIE im sogenannten Mikro-Makro-Paradoxon (u. a. Faust & Leiderer, 2008). Dieses besteht darin, dass auf der Makro-Ebene des statistischen Ländervergleichs mit Aggregatdaten starke Indizien für eine allenfalls eingeschränkte Wirksamkeit der EZ bestehen. Zugleich vermelden Entwicklungsorganisationen auf der Mikro-Ebene konkreter Evaluationen jedoch beachtliche Erfolgsquoten.[5]

So ist auf der Ebene der statistischen Aggregatdatenforschung nach wie vor umstritten, ob EZ Wachstum fördere und Armut vermindere. Gleichzeitig liegen auf dieser Ebene eine Vielzahl an Befunden vor, die nahelegen, dass die Vergabe von EZ-Ressourcen auch politisch und ökonomisch motiviert ist und dass Entwicklungsressourcen zumindest in autoritär regierten Ländern Patronage und autoritäre Strukturen stärken (Hodler & Raschky, 2014; Dutta, Leeson & Williamson, 2013). Im Vergleich zu diesen kritischen Befunden sind Ergebnisse aus Projektevaluierungen deutlich positiver; Erfolgsquoten von über 80% sind die Regel. Aufgelöst wird dieses Mikro-Makro-Paradoxon einerseits durch die Kritik an den Projektevaluationen. Diese seien nicht in hinreichendem Maße unabhängig, in der methodischen und empirischen Qualität unzureichend und zudem meist nicht auf die Identifikation potenzieller, nicht erwünschter negativer Effekte ausgerichtet (Center for Global Development, 2006). Andererseits mehrte sich auch die Kritik an der makro-quantitativen Forschung. Deren Ergebnisse seien geplagt durch oftmals schlechte Datenqualität, notorische Endogenitätsprobleme und ein für praktische Handlungsempfehlungen oftmals zu hohes Abstraktionsniveau.[6] Auch aus dieser Perspektive bieten sich unabhängige und methodisch verlässlichere Wirkungsevaluierungen auf Projekt- und Programmebene als ein relevanter Beitrag zur Auflösung des Paradoxons an.

*Weltinnenpolitik*: Der dritte Katalysator für die zunehmende Bedeutung rigoroser Wirkungsevaluierung lässt sich mit der Metapher der „Weltinnenpolitik" umschreiben, die auf die ansteigenden Interdependenzen zwischen den Staaten und der Verquickung von

Außenbeziehungen und Innenpolitik verweist. In der Entwicklungspolitik manifestieren sich diese Interdependenzen in der wachsenden innergesellschaftlichen Relevanz von Entwicklungsherausforderungen der Entwicklungs- und Schwellenländer in Geberländern. Konflikt- und ökologisch induzierte Migrationsströme, Klimawandel, Terrorismus, ökologische Ressourcenknappheit: Drängende Herausforderungen in Entwicklungs- und Schwellenländern betreffen heute viel unmittelbarer die Gesellschaften der Geber von EZ. Damit einher geht auch ein breiteres Interesse an deren Wirksamkeit zumal die innenpolitischen Auseinandersetzungen mit dem Erstarken des Rechtspopulismus an Intensität gewonnen haben. Kurzum: Die Wirksamkeit der EZ ist nicht mehr nur bürokratisches und akademisches Interessengebiet in einem normativ aufgeladenen, aber politisch etwas randständigen Politikfeld; sie wird angesichts globaler Entwicklungsherausforderungen wie Klimawandel, Gesundheitsschutz, Biodiversität und Gewaltkonflikten zunehmend Gegenstand einer auch innenpolitisch motivierten Debatte.

## 3 Rigorose Wirkungsevaluierung als Instrument der Wirkungsmessung

Die Forderungen nach anspruchsvollen Evaluierungen zur Wirksamkeit der EZ können aus funktionaler Perspektive wie auch aus der Wirksamkeitsdebatte gut abgeleitet werden. Doch stellt sich aus evaluatorischer Perspektive die Frage, inwieweit RIE mit ihren experimentellen und quasi-experimentellen Verfahren der hierfür geeignete Evaluierungstypus sind.

Können Eltern durch die Vergabe konditionierter Haushaltszuschüsse dazu motiviert werden, ihre Töchter in die Schule zu schicken und sollten diese Haushaltszuschüsse eher der Mutter oder dem Vater gegeben werden? Sollte die chronisch anzutreffende Misswirtschaft beim Straßenbau in ruralen Gebieten eher durch lokale Partizipation der Bevölkerung reduziert werden oder stellt eine strenge Kontrolle durch den Rechnungshof eine bessere Alternative dar? Verhindert die Vergabe von Kurzzeitjobs an Bewohner*innen von Flüchtlingslagern das Potenzial von Gewaltkonflikten? Führt die Beratung kommunaler Verwaltungen bei der Raumplanung zu verbessertem Katastrophenschutz und verbesserter Anpassung an den Klimawandel? Hat die Wahlbeobachtung durch internationale Akteure einen Einfluss auf das Wahlergebnis in semidemokratischen Regimen und wird die Opposition hierdurch gestärkt?

All dies sind Fragen nach den Wirkungen von Entwicklungsmaßnahmen, für die es unterschiedliche aber gut begründete Ansichten geben kann. Einen vergleichsweise eindeutigen Befund darüber, ob und wie stark eine (durchschnittliche) Wirkung einer Intervention auf die Zielgruppe eingetreten ist, können zumindest für den Einzelfall rigorose Wirkungsevaluierungen geben. Wenn etwa – wie in der internationalen Wahlbeobachtung üblich – Wahlbeobachter ihre zu beobachtenden Wahldistrikte via Zufallsstichprobe zugeteilt bekommen und ebenso zufallsbasiert eine Kontrollgruppe ohne Wahlbeobachter gezogen wird, dann kann der Einfluss der Beobachter auf das offizielle Wahlergebnis in den Distrikten ermittelt werden (Hyde, 2007). Ähnliches kann mit Blick auf lokale Infrastrukturvorhaben geschehen. Gemäß einer Zufallsziehung können unterschiedliche Gruppen mit Dörfern gebildet werden, in denen lokale Stra-

ßenbauprojekte geplant sind: Solche, in der parallel zur Maßnahme lokale Partizipationsmechanismen der sozialen Kontrolle eingeführt werden, solche, in der eine Untersuchung durch den Rechnungshof angekündigt wird, und eine Kontrollgruppe, also eine Gruppe von Dörfern, in denen keine der Maßnahmen eingeführt werden. Sind die Gruppengrößen hinreichend groß, so kann der durchschnittliche Qualitätsunterschied der gebauten Straßen, deren Materialgüte und deren Kosten auf die unterschiedlichen Interventionen zurückgeführt werden (Olken, 2007).

Auch quasi-experimentelle Designs können bei überzeugenden Identifikationsstrategien einen ähnlich verlässlichen Effekt einer Intervention wie randomisierte Studien identifizieren, wenn die Randomisierung etwa aus administrativ-politischen, zeitlichen, ethischen oder finanziellen Gründen nicht möglich ist. So können in einer Vielzahl von Fällen etwa Matching-Verfahren eingesetzt werden, bei denen nach entsprechenden Erhebungen Interventions- und Kontrollgruppe verglichen und mittels statistischer Verfahren, verzerrende Störgrößen eliminiert werden: Sei es bei der Untersuchung von Reformberatung auf lokaler Ebene (Leppert, Hohfeld, Lech & Wencker, 2018), der Untersuchung von konditionierten Haushaltszuschüssen an bedürftige Familien (Rawlings & Rubio, 2005; Molina, Millán, Barham, Macours, Maluccio & Stampini, 2019) oder auch der Entsendung junger Freiwilliger in Entwicklungsländer (Polak, Guffler & Scheinert, 2018).

Der entscheidende Vorteil experimenteller und quasi-experimenteller Verfahren ist dabei, dass sie den Einfluss typischer Störgrößen eliminieren. Der einfache Vorher-Nachher-Vergleich kann den potenziellen Effekt einer Maßnahme in den vielfach besonders dynamischen Kontexten der EZ kaum identifizieren, wenn auf Zielindikatoren viele inkonstante Drittvariablen wirken. Ebenso schwerwiegend sind gerade in der EZ oftmals anzutreffende Selektionsverzerrungen. Ist es tatsächlich die Trainingsmaßnahme, die Erfolge bei der Zielgruppe hervorruft oder haben sich nicht gerade besonders motivierte und qualifizierte Akteure in die Interventionsgruppe hineinselektiert? Auf aggregierter Ebene wählen Akteure der EZ oftmals besonders leistungsfähige oder auch besonders marginalisierte oder bedürftige räumliche Einheiten oder Organisationen wie Schulen, Gemeinden oder Provinzverwaltungen für ihre Maßnahmen aus. Dieser Bias erfolgt oft, weil sich durch die Wahl qualifizierter Einheiten eine höhere Erfolgswahrscheinlichkeit der Intervention erhofft wird oder auch weil die Wahl besonders prekärer Verhältnisse für die Intervention aus normativen Gründen geboten erscheint. Zur Folge hat eine solch systematische Auswahl jedoch schwerwiegende Endogenitätsprobleme bei der Wirkungsmessung. Der Vorteil experimenteller und quasi-experimenteller Designs besteht daher in der Eliminierung dieser Störgrößen und einer vergleichsweise eindeutigen Zuordnung von Ursache und Wirkung.

Doch sind RIE mit ihren experimentellen und quasi-experimentellen Komponenten deshalb Goldstandard der Evaluation oder gar methodischer Königsweg der Sozialwissenschaften und wie weit tragen die Argumente, die Kritiker von RIE gemeinhin anführen? Die am häufigsten betrachteten Kritikpunkte lassen sich in unterschiedliche Kategorien einordnen. Es sind erstens Kritikpunkte, die sich auf das theoretisch-analytische Fundament von RIE beziehen und diesen die Vernachlässigung von Theoriebasiertheit und mangelnde Untersuchung von Kausalmechanismen vorwerfen (Deaton, 2010; Deaton & Cartwright, 2018; Donovan, 2018). Es sind zum zweiten normative Kritikpunkte, die den Fokus auf die ethische Unangemessenheit bzw. die Instrumentalisierung von RIE für eine Ökonomisierung der EZ und deren Fokus auf die Kontrollfunktion von Evaluierung legen (Baele, 2013; Donovan, 2018; Bédécarrats,

Guérin & Roubaud, 2019). Drittens schließlich sind es Kritikpunkte, die sich auf Aggregationsfähigkeit und die thematische Reichweite von RIE beziehen (Deaton, 2010; Pritchett & Sandefur, 2015; Kvangraven, 2019).

*Ad 1) theoretische Fundierung und Kausalmechanismen:* Eine oftmals anzutreffende Kritik an RIE führt an, dass experimentelle und quasi-experimentelle Designs nicht oder nur geringfügig theoriegeleitet und zudem lediglich auf die Identifikation von Kausalzusammenhängen ausgerichtet seien, nicht jedoch auf die der Kausalität zu Grunde liegenden Mechanismen (u. a. Deaton, 2010; Deaton & Cartwright, 2018). Diese Kritik kann fallspezifisch zutreffen, richtet sich aber nur auf sehr rudimentäre experimentelle bzw. quasi-experimentelle Untersuchungen. Qualitativ hochwertige RIE, so der mittlerweile breite Konsens auch unter den Befürworter*innen des Ansatzes, sind theoriegeleitete Untersuchungen, die eine theoriebasierte und hypothesengestützte Interventionslogik eines Programms nutzen, erarbeiten bzw. rekonstruieren und diese auch mit der relevanten breiteren sozialwissenschaftlichen Auseinandersetzung verknüpfen (White, 2009). Dass Theoretiker*innen und Empiriker*innen nicht einfach zusammenfinden (Deaton, 2010) und die theoretische Anleitung und Einbettung empirischer Studien kein einfaches Unterfangen ist, gilt denn auch nicht nur für RIE. Insofern ist ein sporadisch auftretendes Missverhältnis von Theorie und Empirie denn auch kein spezifisches Defizit von RIE, sondern eine Herausforderung für qualitativ hochwertige Evaluation bzw. Sozialforschung insgesamt.

Weiterhin muss sich die Theoriebasiertheit von RIE nicht nur auf die Identifizierung von Kausalzusammenhängen beziehen, sondern kann auch Fragen nach den Ursachen der identifizierten Zusammenhänge nachgehen. Sich der Öffnung solch schwarzer Boxen zu widmen, kann nicht nur mittels quantitativer Mediatorenanalyse geschehen. Vielmehr sind hierfür auch qualitative Komponenten des *Process Tracing* geeignet, die sowohl vor oder auch nach der experimentellen/quasi-experimentellen Erhebung eingebaut werden und auch ethnographische Elemente enthalten können. Hierbei wird aus der aktuellen Diskussion um die Untersuchung von Kausalmechanismen in der Evaluation (Schmitt, 2020) deutlich, dass mittlerweile auch RIE eine Vielzahl von Möglichkeiten nutzen, quantitative und qualitative Komponenten der Datenerhebung miteinander zu kombinieren (Oakley, Strange, Bonell, Allen & Stephenson, 2006; White, 2013; Bell & Peck, 2016; Jimenez et al., 2018).

> „RCTs can play a role in building scientific knowledge and useful predictions but they can only do so as part of a cumulative program, combining with other methods, including conceptual and theoretical development, to discover not ‘what works’, but ‘why things work’.” (Deaton & Cartwright, 2018, p. 2)

Sicherlich ist mit dem Bedeutungsaufschwung von RIE und der Verbreitung experimenteller und quasi-experimenteller Verfahren auch eine zunehmende Zahl solcher Untersuchungen einhergegangen, die weder eine sonderlich ausgeprägte Theoriebasiertheit noch zufriedenstellende Erklärungsangebote hinsichtlich der wirkenden Kausalmechanismen erkennen lassen (Deaton, 2010). Eine grundsätzliche Kritik am Instrument der RIE lässt sich hierdurch jedoch nicht ableiten, da es sich theoriebasiert anwenden lässt und qualitative Komponenten zur Identifizierung von Kausalmechanismen integriert werden können. Insofern hatten die genannten Kritikpunkte aufgrund ihrer fallspezifischen Berechtigung durchaus die positive Auswirkung, Multi-Methoden-Ansätze und Theoriebasiertheit bei der Durchführung von RIE zu befördern.

*Ad 2) Ethische Unangemessenheit, Ökonomisierung und Fokussierung auf die Kontrollfunktion*: Weitere Kritikpunkte sind eher normativen Ursprungs. Hierbei sind zunächst die ethischen Probleme zu nennen, die bei einer Randomisierung von Entwicklungsinterventionen entstehen können. Vor allem wenn Interventionen das entwicklungspolitische *Do-No-Harm*-Gebot zu verletzen drohen und mit der Randomisierung negative Konsequenzen für die Interventions- oder Kontrollgruppe verbunden sind, sollten Feldexperimente aus normativ-ethischen Gründen unterbleiben (u. a. Barret & Carter, 2010; Baele, 2013). Es ist daher einerseits notwendig, die Durchführung eines geplanten Experiments vor Beginn durch eine unabhängige Ethikkommission bewerten zu lassen.

Andererseits droht die intensive Diskussion um die ethischen Standards von Feldexperimenten in der EZ in eine Schieflage zu geraten. Die makro-quantitative Wirksamkeitsdebatte hat nicht unerhebliche Zweifel an der entwicklungspolitischen Wirksamkeit genährt und viele Projektevaluierungen lassen Bedenken aufkommen, ob deren Wirksamkeitsbefunde methodisch hinreichend abgesichert sind (u. a. Noltze, Euler & Verspohl, 2018). Angesichts dieser Ausgangssituation ist es verwunderlich, dass an nicht randomisierten Interventionen bislang nur wenig Kritik an fehlenden ethischen Standards aufgekommen ist. Bislang wird das Gros an Maßnahmen mit ihren potenziellen Konsequenzen auf etwa Steuer-, Wahl-, Bildungs- oder Konfliktverhalten ohne größere ethische Prüfungen akzeptiert, obwohl über die zu erwartenden Wirkungen Unsicherheit herrscht. Angesichts dieser Ausgangslage scheint die auch ethisch begründbare Forderung nach mehr RIE zwecks Reduzierung dieser Unsicherheit nachvollziehbar (Center for Global Development, 2006). Hierbei kann zudem darauf verwiesen werden, dass sich die ethische Kritik an RIE fast ausnahmslos auf die Randomisierung in Feldexperimenten beschränkt. Quasi-experimentelle Verfahren hingegen bieten aus ethischer Perspektive ein breiteres Einsatzfeld und können wohl begründete ethische Bedenken bei einzelnen Feldexperimenten insofern oftmals gut ersetzen.

Ein zweiter normativer Kritikpunkt postuliert, dass RIE vornehmlich die Kontrollfunktion von Evaluierung zu bedienen suchen und einer Ökonomisierung des Politikfeldes bzw. dessen Anpassung an neoliberale Ideen Vorschub leisten (Bédécarrats Guérin & Roubaud, 2019; Kvangraven, 2019). Allerdings haben beide Argumente aus einer evaluatorisch-analytischen Perspektive wenig Substanz. Sicherlich können RIE auch der Funktion der Rechenschaftslegung dienen, indem sie in Form einer Ex-Post-Evaluierung die Wirkungen einer Entwicklungsmaßnahme erfassen und damit auch als kritische Evidenz über deren Fortführung oder Replizierung herangezogen werden. Gleichwohl sind gerade Feldexperimente – insbesondere solche mit mehreren Interventionsarmen – oftmals so angelegt, dass sie als Begleitforschung großer Sozialprogramme auf das Lernen zu Beginn der Implementierung ausgerichtet sind. So kann die mehrjährig andauernde Ausbreitung und Anpassung großer Sozialprogramme von den Ergebnissen begleitender RIE profitieren. Ebenso sind auch summative RIE nicht nur auf die Kontrollfunktion von Evaluierung ausgelegt, sondern deren Evidenz kann für die Weiterentwicklung und Anpassung der Instrumente dienlich sein. Es mag daher zwar zutreffen, dass RIE oftmals auch aus einer *top-down*-Logik als Kontrollinstrument eingesetzt wurden und nicht auf strukturiertes Lernen und Wissensmanagement ausgerichtet waren (Pritchett, Samji & Hammer, 2013). Gleichwohl zeigen eine Vielzahl von RIE auch, wie diese zu Lern- und Anpassungszwecken genutzt werden können.

Ähnliches gilt prinzipiell auch für den Vorwurf neoliberaler Ökonomisierung der EZ mittels des Instruments von RIE. Es ist möglich, dass kritische Elemente aus RIE in der politischen Auseinandersetzung von Gegnern der EZ aufgenommen und ins Feld geführt werden. Ebenso ist es möglich, dass die Ergebnisse aus RIE für Argumente genutzt werden, Anpassungen und Effizienzsteigerungen zu fordern. In Abhängigkeit von der produzierten Evidenz, können die Ergebnisse aus RIE jedoch auch als Plädoyer für Forderungen wirken, die gegen libertäre Politiken gerichtet sind und etwa auf eine Erhöhung von EZ-Zahlungen zielen. So sind maßgebliche Vertreter*innen von RIE unter den Entwicklungsökonom*innen weder als Befürworter*innen neoliberaler Strategien aufgetreten, noch haben sie für eine Abschaffung der EZ plädiert. Was jedoch von den Verfechter*innen der rigorosen Evidenzagenda zugestanden wird, ist, dass diese Agenda eng verwandt ist mit Kernkonzepten des *New Public Managements*, die den öffentlichen Sektor stärker in Analogie zu marktwirtschaftlichen, individualisierten Anreizsystemen gestalten wollen (White, 2019). Ob die Ausgestaltung von *New-Public-Management*-Systemen jedoch ausschließlich von einer libertären Idee von Marktwirtschaft getrieben wird, ist umstritten.

*Ad 3) Externe Validität und thematische Reichweite:* Ein letztes Bündel an Kritikpunkten befasst sich mit der Reichweite und Aggregationsmöglichkeiten von RIE und offenbart im Unterschied zu den vorherigen Kritikpunkten strukturelle Begrenzungen rigoroser Wirkungsevaluierung. Zunächst ist hierbei das Problem geringer externer Validität zu nennen. Die berechtigte Kritik lautet hierbei, dass die Ergebnisse von RIE letztlich stark kontextgebundene Fallstudien mit einer hohen internen aber geringen externen Validität seien, aus denen eben keine Empfehlungen für andere Kontexte bzw. Populationen gezogen werden können (Prichett & Sandefur, 2015). Gerade für die Ableitung von kontextübergreifenden Empfehlungen ist jedoch ein hohes Maß an externer Validität von großer Bedeutung. Dani Rodrik (2009) beschreibt treffend, dass ebenso wie eine herkömmlich quantitative Studie zusätzliche Argumente außerhalb ihres empirischen Kerns benötige, um ihre interne Validität zu begründen, eine experimentelle Untersuchung zusätzlicher empirischer Argumente zur Begründung ihrer externen Validität bedürfe. Diese Herausforderung kann punktuell durch die Wiederholung des Experiments oder dessen paralleler Durchführung in unterschiedlichen Kontexten behoben werden. Gleichwohl ist der bloße Vergleich der Ergebnisse für eine robuste Aussage über die externe, kontextunabhängige Validität nicht ausreichend.

Hierzu können systematische Reviews genutzt werden, die speziell auf die Aggregation von Ergebnissen aus experimentellen und quasi-experimentellen Studien ausgerichtet sind und mittlerweile wichtiger Bestandteil der von Howard White (2019) postulierten Evidenzagenda in der EZ sind. Systematische Reviews in der sozialwissenschaftlichen Entwicklungsforschung sind ein vergleichsweise neues Phänomen und folgen in ihrer Methodik dabei den Verfahren aus Medizin und Gesundheitsökonomik. Sie finden und bewerten zunächst nach einem ex ante bestimmten Protokoll möglichst alle existierenden Studien, die das erwünschte Entwicklungsziel zum Gegenstand haben (Mallett, Hagen-Zanker, Slater & Duvendack, 2012). Sodann werden die Daten derjenigen Studien aggregiert, die strenge Qualitätskriterien erfüllen, um hierdurch einen durchschnittlichen, kontextübergreifenden Durchschnittseffekt der Intervention und dessen statistische Signifikanz zu ermitteln. Insofern sind Systematische Reviews ein geeignetes Mittel, um die existierende Evidenz aus einzelnen RIE zu einem spezifischen Outcome zusammenzu-

führen und damit das Problem geringer externer Validität zu reduzieren (Mallett, Hagen-Zanker, Slater & Duvendack, 2012; White, 2018). Sie stellen damit eine begrüßenswerte Ergänzung zu einzelnen RIE und makroquantitativen Studien dar.

Gleichwohl werden Systematische Reviews im Unterschied zur Medizin das Problem geringer externer Validität absehbar kaum lösen. Zu gering ist oftmals die Fallzahl an rigorosen Wirkungsevaluierungen, die für einen Systematischen Review in Frage kommen, zu unterschiedlich sind die zu Grunde liegenden Interventionen und zu vielfältig die sozialen Kontexte, in denen keine RIE stattgefunden hat (Mallett, Hagen-Zanker, Slater & Duvendack, 2012). Zudem setzt sich auch bei Vertreter*innen der rigorosen Evidenzagenda dabei zunehmend die Erkenntnis durch, dass theoriegeleitete und um qualitative Methodenelemente ergänzte Systematische Reviews einen Erkenntnismehrwert bringen (Jimenez et al., 2018; White, 2018). Insofern scheint sich zu bestätigen, dass traditionelle, fallübergreifende quantitative Studien mit Herausforderungen interner Validität für die Politikberatung ebenso erkenntnisreich sind wie einzelne rigorose Wirkungsevaluierungen oder auch systematische Reviews mit ihren Beschränkungen externer Validität (Pritchett & Sandefur, 2015). Erneut besteht der Königsweg damit in einer Kombination unterschiedlicher, auf die spezifische Fragestellung zugeschnittener Verfahren aus dem Baukasten der empirisch-analytischen Sozialforschung, um ein hohes Maß an interner und externer Validität zu erzielen.

Hinsichtlich der thematischen Reichweite von RIE hat sich gezeigt, dass diese nicht auf wenige Sektoren beschränkt bleiben müssen. Ursprünglich stark auf die sozialen Sektoren wie Gesundheit, Bildung sowie auf Armutsbekämpfung beschränkt, hat sich in den letzten Jahren gezeigt, dass RIE auch in den allermeisten anderen Sektoren der Entwicklungszusammenarbeit angewendet werden kann. RIE können im Governance-Bereich durchgeführt werden (Funk, Groß, Leininger & Schiller, 2019), sind auch in der (angelsächsischen) Politikwissenschaft mittlerweile verankert (Humphreys & Weinstein 2009; Hyde, 2015) und finden ebenso bei Umwelt- und Klimaanpassungsthemen sowie auch in der humanitären Hilfe Anwendung. Gerade mittels quasi-experimenteller Untersuchungsdesigns und unter Verwendung neuer Datentypen wie etwa Satellitendaten sind auch sektorübergreifende bzw. multisektorale Evaluierungen möglich (Leppert, Hohfeld, Lech & Wencker, 2018).

Wenn auch deutlich häufiger anwendbar als noch vielfach in der Entwicklungspraxis angenommen, sind RIE jedoch hinsichtlich ihrer Evaluierungsgegenstände strukturell begrenzt, da eine Vielzahl von EZ-Interventionen mit experimentellen- und quasi-experimentellen Methoden kaum zu evaluieren sind. Hierzu zählen insbesondere viele Maßnahmen auf der Meso- oder Makroebene, die sich der Reformberatung und Reformimplementierung auf regionaler oder nationaler Ebene widmen. Die begrenzte Fallzahl – ein oder wenige Ministerien, Behörden, NGOs, etc. – der adressierten Einheiten macht dann ein (quasi-)experimentelles Design unmöglich. Zudem sind nicht alle entwicklungspolitisch relevanten Evaluationsfragen auf Wirkungen gerichtet. Institutionelle Evaluierungen einzelner Organisationen, Prozessevaluationen, Portfolioanalysen und (zumindest teilweise) auch Strategie- und Instrumentenevaluationen sind nicht oder nur begrenzt mit RIE zu bearbeiten. Diese strukturelle Begrenzung des Instruments sollte jedoch nicht dazu führen, RIE „Kleinteiligkeit“ und mangelnde Fähigkeit zur Thematisierung größerer theoretischer und systemischer Fragen zu unterstellen, wie dies einige Kritiker*innen tun (Kvangraven, 2019; Bédécarrats Guérin & Roubaud, 2019). RIE sind in einer Vielzahl von Fällen gut geeignet, die Wirksamkeit kon-

kreter Entwicklungsmaßnahmen zu identifizieren. Sind sie theoriegeleitet, dann lassen sich ihre Befunde in vielen Fällen auch an breitere Debatten in Entwicklungsforschung und Sozialwissenschaften koppeln; so etwa in der Arbeitsmarktforschung oder in der Diskussion um direktdemokratische Partizipationsmechanismen. Nur weil viele RIE auf der Ebene der Individuen ansetzen, müssen sie nicht notwendigerweise einer Verantwortungsverschiebung hin zum Individuum Vorschub leisten, können eine Bewertung von Politiken auf der Makroebene gleichwohl nicht ersetzen.

## 4 Anwendung und Nutzung rigoroser Wirkungsevaluierung in der EZ

Angesicht der Bedeutung von RIE ist neben der Frage nach deren wissenschaftlich-evaluatorischer Eignung zu klären, wie es um die praktische Durchführung und Nutzung von RIE in der internationalen EZ bestellt ist. Ein erster Anhaltspunkt für die Durchführung ist das größte Repositorium für RIE in der angewandten Entwicklungsforschung, das gegenwärtig von der International Initiative for Impact Evaluation (3ie) betrieben wird. Das Repositorium umfasst gegenwärtig über 3.700 einzelne RIE sowie 722 Systematische Reviews (www.3ie.org), die entlang unterschiedlicher Kriterien geordnet werden können. Die tatsächlich existierende Zahl von durchgeführten RIE in der internationalen EZ dürfte allerdings deutlich größer sein und war in den letzten beiden Jahrzehnten durch kontinuierlich hohe Steigungsraten gekennzeichnet (Manning, Goldmann & Hernandez, 2020). Die überwiegende Mehrzahl dieser RIE wird aus Mitteln der EZ (ko-)finanziert; sei es direkt aus den Projekten und Programmen oder durch die Finanzierung aus größeren Fonds oder Förderprogrammen. Insgesamt ist die Finanzierung und Nutzung von RIE in der Entwicklungsforschung daher immer noch stark durch die Geberorganisationen getrieben. Gleichzeitig ist aber auch ein Trend erkennbar, wonach insbesondere in Schwellenländern wie Mexiko, Indien, Kolumbien oder Indonesien die Beauftragung von RIE als summative Evaluierungen oder Begleitforschung durch staatliche Entscheidungsträger zunimmt und hiermit eigene Sozialprogramme analysiert werden (Manning, Goldmann & Hernandez, 2020). Teils werden aber auch diese Initiativen von Organisationen der EZ (ko-)finanziert.

In der internationalen EZ ist die Finanzierung, Anwendung und Nutzung von RIE im internationalen System durch ein hohes Maß an Heterogenität gekennzeichnet und nicht in einen ordnungspolitischen Rahmen eingebettet. Am stärksten ausgeprägt und systematisiert ist die Anwendung und Nutzung von RIE in einigen internationalen Finanzorganisationen und Entwicklungsfonds. Hierbei zählt die Weltbank als Vorreiter, da sie über ein großes Programm zur Durchführung von RIE verfügt (Development Impact Evaluation DIME). In Kooperation mit Partnerländern und ko-finanziert von anderen Gebern, werden dort eine Vielzahl von RIE und Systematischen Reviews durchgeführt und veröffentlicht. Auch in einigen anderen internationalen Entwicklungsbanken und Entwicklungsfonds ist die Anwendung von RIE verbreitet, allerdings in geringerem Umfang und weniger systematisch institutionalisiert als in der Weltbank (u. a. Interamerican Development Bank, 2017).

Parallel zu den Aktivitäten der internationalen Organisationen der EZ gibt es mittlerweile eine Vielzahl internationaler Nichtregierungsorganisationen (NGOs), die sich

als Expertenorganisationen im Schnittfeld zwischen Wissenschaft und Praxis etabliert haben.[7] Organisationen wie das am Massachusetts Institute of Technology (MIT) angesiedelte J-PAL (Abdul Latif Jameel Poverty Action Lab), IPA (Innovations for Povery Action) oder 3ie haben sich dabei auf die Durchführung von RIE, deren Verbreitung und – im Falle von 3ie – auch als Finanzierungsplattform für RIE spezialisiert. Wenngleich ebenfalls in überwiegendem Maße durch Mittel der internationalen EZ finanziert, setzen sich diese Organisationen zum Ziel, die Anwendung und Nutzung von RIE durch Entwicklungs- und Schwellenländer zu fördern. Insbesondere 3ie verfolgt zudem das Ziel, die Evidenz aus RIE transparent aufzubereiten, zu systematisieren und als kostenloses, globales Kollektivgut zur Verfügung zu stellen (siehe www.3ie.org).

Im Unterschied zu den sichtbaren Aktivitäten auf der Ebene der Finanzorganisationen und denjenigen von NGOs im Schnittfeld zwischen Wissenschaft, Evaluierung und Praxis ist die Nutzung von RIE auf bilateraler Ebene unübersichtlich. Vornehmlich unter den angelsächsischen und partiell auch bei den nordischen Gebern werden RIE finanziert und genutzt, sind aber keinesfalls dominantes Instrument der Evaluierung. Trotz der vergleichsweise breiteren Akzeptanz und Nutzung von RIE ist aber auch in Ländern wie den USA oder Großbritannien über Zeit und Sektoren kaum eine systematische Verankerung und Normierung von RIE in der EZ erkennbar (Manning & White, 2014). Eine eindeutige Zuordnung der Evaluierungspraxis nach Geberländern wird auch dadurch erschwert, weil innerhalb dieser verschiedene Durchführungsorganisationen mit unterschiedlichen Evaluierungspraktiken existieren können. In anderen bedeutenden Geberländern wie Deutschland, Frankreich und Japan war die Beauftragung von RIE durch die Akteure der EZ bislang eher die Ausnahme als die Regel. In Deutschland etwa ist während der letzten Dekade die Expertise in Universitäten und anderen Forschungseinrichtungen zu RIE stetig angewachsen und mittlerweile auch jenseits der Entwicklungsökonomie verbreitet. Gleichwohl wurden RIE seitens der finanziellen und technischen Zusammenarbeit bislang nur sporadisch beauftragt, wobei in der deutschen EZ gegenwärtig eine Trendwende hin zu einer breiteren Akzeptanz und Nutzung von RIE erkennbar ist (Bruder, Faust & Krämer, 2019; Florian, Krisch, Till & Hermanns, 2019).

Trotz der international heterogenen Landschaft, was Anwendung und Nutzung von RIE anbelangt, hat sich durch die Dynamik in der Entwicklungsforschung eine praxisorientierte, methodisch wie inhaltlich geprägte epistemische Gemeinschaft (Haas, 1992) mit stark instrumentellem Fokus auf das Policy-Instrument von RIE herausgebildet. Insofern verfügt diese Gemeinschaft auch über Merkmale einer *instrumental constituency* (Simons & Voss, 2018). Sie hat Überlappungen mit der primär universitär orientierten Wissenschaft mit ihrem Fokus auf möglichst anspruchsvollen Feldexperimenten, ist insgesamt aber pragmatischer und versucht als Scharnier zwischen Wissenschaft und Implementierungspraxis eine stärker anwendungsbezogene Nutzung von RIE für praktische Zwecke voranzutreiben. Die Protagonisten dieser epistemischen Gemeinschaft arbeiten denn auch nur teils an Universitäten und Hochschulen, sondern oftmals in Evaluierungseinrichtungen bilateraler und multilateraler EZ-Organisationen, in den genannten wissenschaftsnahen NGOs oder partiell auch in staatlichen Evaluierungs- und Analyseeinheiten von Entwicklungs- und Schwellenländern. Auch ist diese Gemeinschaft in nationalen und internationalen Evaluierungsverbänden vertreten. Doch aufgrund der anhaltenden – auch epistemologischen – Auseinandersetzungen um die Angemessenheit und Anwendbarkeit von RIE sind die Verbände bislang nicht als

Föderatoren von RIE aufgetreten. Ähnliches gilt auch für das Evaluierungsnetzwerk innerhalb der OECD, in dem alle bilateralen Geber der OECD sowie auch multilaterale Organisationen (als Beobachter) vertreten sind. Wenngleich dieses Netzwerk in der Vergangenheit als ein bedeutsamer Standardsetzer in der internationalen Entwicklungsevaluierung aufgetreten ist, waren die Ansichten innerhalb der Gebergemeinschaft zur Anwendung von RIE bislang zu heterogen, als dass dieses Netzwerk eine substanzielle Rolle als Föderator von RIE hätte einnehmen können.

Insofern hat sich die Debatte um RIE bislang als Barriere für eine systematischere und strukturiertere Anwendung von RIE auf internationaler Ebene erwiesen, da sie auf der Ebene von Ministerien, multilateralen Organisationen, bilateralen Durchführungsorganisationen und auch zivilgesellschaftlichen Akteuren ganz unterschiedlich motivierte Vorbehalte provoziert hat; von epistemologischen Auseinandersetzungen zwischen einem empirisch-analytischen und einem konstruktivistisch-postkolonialen Wissenschaftsverständnis bis hin zu Vorbehalten, was Anwendbarkeit und Kontrolle von RIE in der konkreten Praxis anbelangt. Mit einigen relevanten Ausnahmen insbesondere in multilateralen Organisationen, den USA und Großbritannien erfolgte die Anwendung meist dezentralisiert und war oft mehr von wissenschaftlichen Dispositionen und praktischen Erkenntnisinteressen von Projekt- oder Programmleitungen abhängig als von einer gezielten Umsetzung von Evidenzstrategien der Entwicklungsorganisationen. Damit bleibt aber auch der praktische Wert von RIE bislang weit hinter ihrem Potenzial zurück. Um dieses künftig auszuschöpfen, ist es notwendig, dass Entwicklungsorganisationen Strategien, Anreizsysteme und finanzielle Ressourcen für eine situationsangemessene Nutzung von RIE etablieren bzw. bereitstellen sowie die hieraus gewonnene Evidenz systematischer und nicht nur punktuell nutzen und zudem auch der Allgemeinheit im Sinne eines Kollektivgutes zur Orientierung und Aggregation zur Verfügung stellen.

## 5 Fazit

Rigorose Wirkungsevaluierungen mit ihren experimentellen und quasi-experimentellen Komponenten haben in den letzten beiden Dekaden an Bedeutung in der Evaluierung entwicklungspolitischer Maßnahmen gewonnen. Eine größer werdende Zahl an Ökonom*innen und mittlerweile auch Wissenschaftler*innen anderer Sozialwissenschaften sind entsprechend ausgebildet worden und plädieren für eine weitere Verbreitung (quasi-)experimenteller Analysen in der Entwicklungsforschung. Der Vorteil besteht dabei im Vergleich zu anderen Methoden darin, intendierte und nicht-intendierter Wirkungen von Maßnahmen besser identifizieren zu können. In der EZ bedienen RIE daher zum einen den funktional begründbaren, hohen Bedarf an Evaluierung. Zum anderen bedienen sie die Forderung nach zuverlässiger Evidenz über die Wirkungen der EZ aus einer zunehmend kritischen Wirkungsdebatte.

Doch trotz dieser nachvollziehbaren Genese von RIE in der Entwicklungszusammenarbeit hält die Debatte um deren Vor- und Nachteile an. Aus empirisch-analytischer Sicht spricht allerdings wenig gegen eine Verbreitung von RIE in Entwicklungsforschung und Entwicklungszusammenarbeit. Insgesamt haben sich „die Bedenken hinsichtlich grundsätzlicher oder sektoraler Einschränkungen der Anwendbarkeit solcher Verfahren, unüberwindbarer ethischer Probleme oder nicht ausreichend theorieba-

sierter Vorgehensweisen als unbegründet, lösbar oder nur für eingeschränkte Bereiche als relevant erwiesen“ (Bruder, Faust & Krämer, 2019, S. 2). RIE können vielmehr in unterschiedlichen Sektoren theoriebasiert, mit anderen Methoden kombiniert und ethisch vertretbar eingesetzt werden. Sie müssen weder einer neoliberalen Ideologie verpflichtet sein, einen Kontrollbias haben noch kleinteilige akademische Reparaturarbeiten ohne Bezug zu größeren Debatten sein. Oft liegt denn der Ursprung der Kritik auch in den nur begrenzt miteinander versöhnbaren Schulen eines empirisch-analytischen bzw. verstehend-hermeneutischen Wissenschaftsverständnisses.

Gleichwohl sind RIE nicht für alle wirkungsbezogenen Fragestellungen der EZ geeignet und sollten auch nicht präferiertes Verfahren der Evaluation bei jedweder entwicklungspolitischen Maßnahme sein. Schließlich kann auch die Aggregation von RIE in Systematischen Reviews nicht strukturelle Begrenzungen für die externe Validität auflösen. Auch hier ist eine Kombination unterschiedlicher (quantitativer und qualitativer) Methoden die zielführende Variante und sollte stärker genutzt werden. Es gilt somit auch für RIE und die Evaluation der EZ ein zentraler Grundsatz anwendungsorientierter Forschung, wonach die Methode dem praktischen Erkenntnisinteresse zu folgen hat und nicht umgekehrt. RIE sind dann besonders zielführend, wenn die potenziellen Wirkungen einer Intervention im Fokus des Erkenntnisinteresses stehen und diese Intervention viele Akteure in ähnlicher Weise erreichen soll (Bruder, Faust & Krämer, 2019, S. 2). Kurzum: RIE sind weder Goldstandard der Evaluation noch Königsweg der Entwicklungsforschung, können aber sehr wohl breit und erkenntnisgewinnend eingesetzt werden.

Was schließlich die Durchführung und Nutzung in der Praxis der internationalen EZ anbelangt, so ist diese insgesamt über die letzten zwei Dekaden deutlich gestiegen. Gleichzeitig ist die Uneinheitlichkeit und Ungleichzeitigkeit der Anwendung, Verbreitung und Nutzung von RIE weiterhin gängige Praxis in der internationalen EZ. Obwohl sich mittlerweile eine stark praxis- und anwendungsorientierte epistemische RIE-Gemeinschaft herausgebildet hat, gibt es bis auf wenige Ausnahmen kaum organisationsweite und schon gar nicht organisationsübergreifende Strategien und Anreizsysteme zur Anwendung und Nutzung von RIE für übergreifendes Lernen und Rechenschaftslegung. Es liegt dabei nahe, dass die intensive Auseinandersetzung um die analytische und normative Eignung sowie die Nutzungsmöglichkeiten von RIE ein Hindernis für die strukturiertere Anwendung und Nutzung von RIE auf internationaler Ebene gewesen ist. Allerdings ist die Heterogenität mit Blick auf die Anwendung und Nutzung von RIE auf internationaler Ebene nur begrenzt eine Besonderheit von RIE. Denn trotz des im Vergleich zu anderen Politikfeldern erreichten hohen Maßes an Austausch und Standardisierung in der Entwicklungsevaluierung ist die Heterogenität der Anwendung und Nutzung von RIE auch ein Strukturmerkmal eines allenfalls begrenzt funktionalen Wissensmanagements der internationalen EZ.

## Anmerkungen

1 Für wertvolle Anregungen zu einer früheren Version dieses Beitrags danke ich Martin Bruder, Marion Krämer, Julia Leininger, Elisabeth Schneider, Holger Straßheim sowie den Gutachter*innen dieses Beitrags.

2 Wie auch in experimentellen Verfahren in Medizin, Gesundheitsökonomie oder Erziehungswissenschaften basieren entwicklungspolitische Feldexperimente (Randomized Controlled Trials - RCTs) auf

der zufallsbasierten Auswahl von Interventions- und Kontrollgruppe. Dabei wird eine Entwicklungsmaßnahme bei einer Gruppe zufällig aus einer Grundgesamtheit gewählter Akteure durchgeführt (Interventionsgruppe) und eine ebenso zufällig gewählte Gruppe dient als Kontrollgruppe. Bei hinreichender Größe können beide Gruppen in ihren beobachtbaren wie nicht beobachtbaren Eigenschaften als gleich erwartet werden. Die Differenz der Zielgröße zwischen beiden Gruppen nach der Durchführung der Maßnahme schätzt einen unverzerrten Nettoeffekt der Intervention. Neben RCTs können auch quasiexperimentelle Verfahren ähnlich verlässlich sein. Diese Verfahren führen zwar kein Experiment zu Beginn einer Intervention durch, bedienen sich aber bestimmter statistischer Verfahren, um hierdurch hinreichend große Vergleichsgruppen zu bilden und Störgrößen systematisch ausschließen (Bruder, Faust & Krämer, 2019, S. 1; einführend siehe u. a. Caspari & Barbu, 2008). Für eine anspruchsvolle Einführung in die Methoden und Anwendungsfelder rigoroser Wirkungsevaluierung siehe Frölich & Sperling, 2018.

3 Aus der umfangreichen kritischen Befassung mit RIE siehe u. a. Acemoglu 2009; Deaton, 2010; Barrett & Carter, 2010; Pritchett, Samji & Hammer, 2013; Kvangraven, 2019; Donovan, 2018; Bédécarrats, Guérin & Roubaud, 2019.

4 Zu den Kenntnissen der deutschen Bevölkerung über die Entwicklungszusammenarbeit und die Nachhaltigen Entwicklungsziele der Agenda 2030 siehe Schneider & Gleser, 2018.

5 So verweist beispielsweise der letzte Evaluierungsbericht der KfW aus dem Jahr 2019 auf eine Erfolgsquote von 77% im Zeitraum von 2017-2018 (KfW Entwicklungsbank, 2019, S. 64). Die Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) gibt in ihrem letzten Evaluierungsbericht von 2017 an, dass auf einer fünfstufigen Bewertungsskala 98% der im Berichtszeitraum (2016-2017) evaluierten Maßnahmen mit der Note „eher erfolgreich“ oder besser bewertet wurden (GIZ, 2018, S. 16). Auch im internationalen Bereich werden gemeinhin Erfolgsquoten von mehr als 75% berichtet.

6 Endogenitätsprobleme bestehen insbesondere dann, wenn aus einer identifizierten Korrelation zweier Variablen nicht auf einen kausalen Zusammenhang zwischen beiden Variablen geschlossen werden kann, weil nicht plausibel auszuschließen ist, dass eine (unbeobachtete) dritte Variable beide gemeinsam beeinflusst oder die Kausalität auch in die umgekehrte Richtung verläuft.

7 Neben den großen internationalen Akteuren wie DIME, J-Pal, IPA und 3ie existiert eine Vielzahl weiterer Organisationen wie GRADE (Grupo de Analisis para el Desarrollo), CEGA (Center for Effective Global Action) oder C4ED (Center for Evaluation and Development). Letztere haben sich zumeist auf die Durchführung von RIE im Auftrag internationaler Entwicklungsorganisationen spezialisiert und haben ihren Sitz sowohl in den OECD-Ländern als auch in Entwicklungsländern. Schließlich ist noch die auf die Durchführung systematischer Reviews spezialisierte Campbell Collaboration zu nennen, die in den letzten Jahren ihre Aktivitäten auf die Entwicklungsforschung ausgeweitet hat (White, 2019).

## Literatur


Acemoglu, Daron (2009). Theory, General Equilibrium, Political Economy and Empirics in Development Economics. *Journal of Economic Perspectives, 24* (3), 17-32.

Baele, Stéphane J. (2013). The ethics of New Development Economics: is the Experimental Approach to Development Economics morally wrong? *The Journal of Philosophical Economics 7* (1), 2-42.

Banerjee, Abhijit V. & Duflo, Esther (2011). *Poor Economics: rethinking poverty & the ways to end it.* Random House.

Barrett, Christopher & Carter, Michael R. (2010). The Power and Pitfalls of Experiments in Development Economics: Some Non-random Reflections. *Applied Economic Perspectives and Policy, 32* (4), 515-548.

Bédécarrats, Florent, Guérin, Isabelle & Roubaud, Francois (2019). All that glitters is not gold. The political economy of randomized evaluations in development. *Development & Change, 50* (3), 735-762.

Bell, Stephen & Peck, Laura (2016). On the “How” of Social Experiments: Experimental Designs for Getting Inside the Black Box. *New Directions for Evaluation, 152*, 97-107.

Bruder, Martin, Faust, Jörg & Krämer, Marion (2019). *Rigorose Wirkungsevaluierung in der deutschen Entwicklungszusammenarbeit*. Policy-Brief 05/2019. Bonn: Deutsches Evaluierungsinstitut der Entwicklungszusammenarbeit.

Caspari, Alexandra & Barbu, Ragnhild (2008). *Wirkungsevaluierungen: Zum Stand der internationalen Diskussion und dessen Relevanz für Evaluierungen der deutschen Entwicklungszusammenarbeit*. Evaluation Working Papers. Bonn: Bundesministerium für wirtschaftliche Zusammenarbeit und Entwicklung.

Center for Global Development (2006). *When Will We ever learn? Improving Lives through Impact Evaluation.* Washington.

Deaton, Angus (2010). Instruments, randomization, and learning about development. *Journal of Economic Literature, 48* (2), 424-55.

Deaton, Angus & Cartwright, Nancy (2018). Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine, 210*, 2-21.

Donovan, Kevin P. (2018). The rise of the randomistas: on the experimental turn in international aid, *Economy and Society, 47* (1), 27-58.

Dreher, Axel, Fuchs, Andreas, Lang, Valentin & Langlotz, Sarah (2018). Migration: Schaffen wir das? *Frankfurter Allgemeine Zeitung* vom 17.09.2018.

Dreher, Axel, Sturm, Jan Egbert & Vreeland, James (2009). Global horse trading: IMF loans for votes in the United Nations Security Council. *European Economic Review, 53* (7), 742-757.

Dutta, Nabamita, Leeson, Peter & Williamson, Cornelia (2013). The Amplification Effect: Foreign Aid's Impact on Political Institutions. *Kyklos, 66* (2), 208-228.

Easterly, William (2002). The cartel of good intentions: the problem of bureaucracy in foreign aid. *The Journal of Policy Reform, 5* (4), 223-250.

Easterly, William (2006). *The White Man's Burden. Why the West's Efforts to Aid the Rest Have Done So Much Ill and So Little Good*. New York.

Faust, Jörg & Leiderer, Stefan (2008). Zur Effektivität und politischen Ökonomie der Entwicklungszusammenarbeit. *Politische Vierteljahresschrift, 49* (1), 129-152.

Faust, Jörg & Michaelowa, Katherina (Hrsg.) (2013*). Politische Ökonomie der Entwicklungszusammenarbeit.* Nomos: Baden-Baden.

Faust, Jörg & Ziaja, Sebastian (2012). *German aid allocation and partner country selection: development-orientation, self-interests and path dependency*. Discussion Paper 7/2012. Bonn: German Development Institute / Deutsches Institut für Entwicklungspolitik (DIE).

Florian, Michael, Krisch, Franziska, Till, Tatjana & Hermanns, Sophie (2019). *Rigorous Impact Evaluation. A Corporate Strategic Review of Causal Analysis during Implementation*. Bonn: Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) GmbH.

Frölich, Markus & Sperlich, Stefan (2018). *Impact Evaluation*. CUP.

Funk, Evelyn, Groß, Lisa, Leininger, Julia & Schiller, Armin (2019). *Erkenntnisse aus der wirkungsorientierten Begleitforschung: Potential und Grenzen der rigorosen Wirkungsanalyse von Governance-Programmen.* Discussion Paper 13/2019. Bonn: Deutsches Institut für Entwicklungspolitik.

GIZ Gesellschaft für Internationale Zusammenarbeit (2018). *Evaluierungsbericht 2017 – Wissen was wirkt*. Eschborn.

Hodler, Roland & Raschky, Paul (2014). Regional Favoratism. *The Quarterly Journal of Economics, 129* (2), 995-1033.

Haas, Peter M. (1992). Introduction: Epistemic Communities and International Policy Coordination. *International Organization, 46 (2)*, 1-35,

Humphreys, Macartan & Weinstein, Jeremy (2009). Field experiments and the political economy of development. *Annual Review of Political Science, 12*, 367-376.

Hyde, Susan (2007). The Observer Effect in International Politics: Evidence from a Natural Experiment. *World Politics, 60* (1), 37-63.

Hyde, Susan (2015). Experiments in International Relations: Lab, Survey, and Field. *Annual Review of Political Science, 18*, 403-424.

Interamerican Development Bank (2017). *IDB's Impact Evaluations – Production Use and Influence*. Washington: Office of Evaluation and Oversight.

Jimenez, Emmanuel, Waddington, Hugh, Goel, Neeta, Prost, Audrey, Pullin, Andrew, White, Howard, Lahiri, Shaon & Narain, Anmol (2018). Mixing and matching: using qualitative methods to

improve quantitative impact evaluations (IEs) and systematic reviews (SRs) of development outcomes. *Journal of Development Effectiveness, 10* (4), 400-421.

KfW Entwicklungsbank (2019). *15. Evaluierungsbericht 2017-2018 – Zu größerer Wirkung in kleineren Städten.* Frankfurt: Kreditanstalt für Wiederaufbau.

Kvangraven, Ingrid H. (2019). *Impoverished economics? Unpacking the economics Nobel Prize*. Open Democracy Net, 18.10.2019.

Leppert, Gerald, Hohfeld, Lena, Lech, Lena & Wencker, Thomas (2018). *Impact, Diffusion and Scaling-Up of a Comprehensive Land-Use Planning Approach in the Philippines. From Development Cooperation to National Policies*. Bonn: German Institute for Development Evaluation (DEval).

Mallett, Richard, Hagen-Zanker, Jessica, Slater, Rachel & Duvendack, Maren (2012). The benefits and challenges of using systematic reviews in international development research. *Journal of Development Effectiveness, 4* (3), 445-455

Manning, Richard, Goldman, Ian & Hernandez Licona, Gonzalo (2020). *The impact of impact evaluation: Are impact evaluation and impact evaluation synthesis contributing to evidence generation and use in low- and middle-income countries?*. WIDER Working Paper 2020/20. Helsinki: UNU-WIDER.

Manning, Richard & White, Howard (2014). Measuring results in development: The role of impact evaluation in agency-wide performance measurement systems. *Journal of Development Effectiveness, 6* (4), 337-349.

Martens, Bertin, Mummert, Uwe & Murrell, Peter (2002). *The Institutional Economics of Foreign Aid*. CUP.

Meyer, Wolfgang, Bär, Thomas, Faust, Jörg, Jan, Susanne v., Silvestrini, Stefan & Wein, Stefanie (2019). Die DeGEval-Standards in der deutschen bilateralen Entwicklungszusammenarbeit. In: Jan Ulrich Hense, Wolfgang Böttcher, Michael Kalman & Wolfgang Meyer (Hrsg.), *Evaluation: Standards in unterschiedlichen Handlungsfeldern - Einheitliche Qualitätsansprüche trotz heterogener Praxis?* (S. 165-182). Münster: Waxmann.

Molina Millán, Teresa, Barham, Tania, Macours, Karen, Maluccio, John A. & Stampini, Marco (2019). Long-Term Impacts of Conditional Cash Transfers: Review of the Evidence. *The World Bank Research Observer, 34* (1), 119-159.

Morgenthau, Hans (1962). A Political Theory of Foreign. *American Political Science Review, 56* (2), 301-309.

Noltze, Martin, Euler, Michael & Verspohl, Ida (2018). *Meta-Evaluierung von Nachhaltigkeit in der deutschen Entwicklungszusammenarbeit.* Bonn: Deutsches Evaluierungsinstitut der Entwicklungszusammenarbeit (DEval).

Oakley, Ann, Strange, Vicki, Bonell, Chris, Allen, Elisabeth & Stephenson, Judith (2006). Process evaluation in randomised control trials of complex interventions. *British Medical Journal, 332*, 413-416.

Olken, Benjamin A. (2007). Monitoring corruption: evidence from a field experiment in Indonesia. *Journal of Political Economy, 115* (2), 200-249.

Organisation für wirtschaftliche Zusammenarbeit und Entwicklung (OECD) (2016). *Evaluation Systems in Development Co-operation – A Review*. Paris.

Polak, Jan T., Guffler, Kerstin & Scheinert, Lara (2017). *Weltwärts-Freiwillige und ihr Engagement in Deutschland*. Bonn: Deutsches Evaluierungsinstitut der Entwicklungszusammenarbeit (DEval).

Pritchett, Lant, Samji, Salimah & Hammer, Jeffrey (2013). *It's All About MeE: Using Structured Experiential Learning ("e") to Crawl the Design Space*. Working Paper 322. Washington: Center for Global Development.

Pritchett, Lant & Sandefur, Justin (2015). Learning from Experiments When Context Matters. *American Economic Review, 105* (5), 471-75.

Rawlings, Laura B. & Rubio, Gloria M. (2005). Evaluating the Impact of Conditional Cash Transfer Programs. *The World Bank Research Observer, 20* (1), 29-55.

Rodrik, Dani (2009). *The New Development Economics: We Shall Experiment, but How Shall We Learn?*. Harvard Kennedy School Working Paper No. RWP08-055.

Schmitt, Johannes (2020). *The Causal Mechanism Claim in Evaluation: Does the Prophecy fulfill?* (mimeo).

Schneider Sebastian H. & Gleser, Solveig H. (2018). *Meinungsmonitor Entwicklungspolitik 2018: Einstellungen zu Entwicklungszusammenarbeit und nachhaltiger Entwicklung*. Bonn: Deutsches Evaluierungsinstitut der Entwicklungszusammenarbeit (DEval).

Simons, Arno & Voss, Jan-Peter (2018). Instrument Constituencies: The politics of policy formulation. *Policy and Society, 37 (1)*, 14-35.

Stockmann, Reinhardt (2004). *Was ist eine gute Evaluation? Einführung zu Funktionen und Methoden von Evaluationsverfahren.* Centrum für Evaluation CEval Arbeitspapier Nr. 9. Saarbrücken.

White, Howard (2009). Theory-Based Impact Evaluation: Principles and Practice. *Journal of Development Effectiveness, 3* (1), 271-284.

White, Howard (2013). The use of mixed methods in randomized control trials. *New directions for evaluations, 138*, 61-73.

White, Howard (2018). Theory-based systematic reviews. *Journal of Development Effectiveness, 10* (1), 17-38.

White, Howard (2019). The twenty-first century experimenting society: the four waves of the evidence revolution. *Palgrave Communications, 5* (47). Online verfügbar unter: https://www.nature.com/articles/s41599-019-0253-6 [30.9.2019].

## *Anschrift des Autors:*

Prof. Dr. Jörg Faust, DEval Deutsches Evaluierungsinstitut der Entwicklungszusammenarbeit, Fritz-Schäffer-Str. 26, 53113 Bonn, E-Mail: joerg.faust@deval.org.